On Natural Variation in Grades in Higher Education, and Its Implications for Assessing Effectiveness of Educational Innovations

Boevé, Anja J.; Meijer, Rob R.; Beldhuis, Hans J. A.; Bosker, Roel J.; Albers, Casper J.

Published in:
Educational Measurement: Issues and Practice

DOI:
10.1111/emip.12283

Publication date:
2019

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2019, Vol. 38, No. 4, pp. 55–66

On Natural Variation in Grades in Higher Education, and Its Implications for Assessing Effectiveness of Educational Innovations

Anja J. Boevé, *University of Groningen and Vrije Universiteit Amsterdam*, Rob R. Meijer, Hans J. A. Beldhuis, Roel J. Bosker, and Casper J. Albers, *University of Groningen*

To investigate the effect of innovations in the teaching–learning environment, researchers often compare study results from different cohorts across years. However, variance in scores can be attributed to both random fluctuation and systematic changes due to the innovation, complicating cohort comparisons. In the present study, we illustrate how using information about the variation in course grades over time can help researchers and practitioners better compare the grades and pass rates of different cohorts of students. To this end, all 375,093 grades from all 40,087 first-year students at a Dutch university during a period of six consecutive years were examined. Overall, about 17% of the variation in grades could be attributed to random variation between years and courses. With respect to passing courses, this percentage was almost 40%. Nonsignificant improvements in grades could be flagged as highly significant when this is ignored, thus leading to an overrepresentation of significant effects in educational literature. As a consequence, too many educational innovations are claimed to be effective.

Keywords: educational innovations, effectiveness, pass rates, reporting bias, statistical modeling

Due to increasing performance-based accountability systems in higher education (Alexander, 2000; Liu, 2011), universities have to keep track of student performance as one of the many indicators of quality and effectiveness. To achieve this, teachers need to demonstrate that the results of student evaluations are taken seriously, and to show how changes, when necessary, improve the teaching and learning environment. As a result, courses are evaluated every year and teachers keep track of how different cohorts of students perform in subsequent years. At the same time, teachers also need to evaluate the success of implemented changes or educational innovations, where an important criterion is often the extent to which student performance has improved. This is difficult to measure in practice, however, because variation in test scores across different years may be due to different factors, including differences in exam difficulty, all sorts of cohort differences, and the effect of educational innovations. A fully experimental design (or randomized controlled trial [RCT]) to study the causal effects of an educational innovation is usually unfeasible in practice, and alternative designs are needed (Carey & Stiles, 2015; West et al., 2008). Thus, comparing course results across years is possible, but this is not an easy task.

To disentangle different sources of variation in this context, the aim of this study was to gain insight into the amount of variation in course grades and pass rates between years across different courses. These variations constitute "naturally expected variability", that is, variability that is not due to the intervention of interest. It differs per research question what should be classified as "naturally expected variability." For instance, when a lecturer wants to study whether a change in the literature that has to be studied has an effect on the course pass rate, she compares the pass rates before and after the change. The naturally expected variability will consist of factors such as cohort composition (e.g., percentage of female students and past performance of the cohort) and the quality of the exam questions. In case of other changes not due to the change in literature, for example, a considerable improvement in information and communication technologies (ICT) facilities in the lecture hall, these circumstances may also contribute to the naturally expected variation. In larger scale studies, there might be more variables that contribute to the natural variation, for example, differences in the quality of the lecturers.

Another layer of natural variability could occur due to external changes to the system. For example, Albers, Vermue, de Wolff, and Beldhuis (2018) studied the effect of a change in

Anja J. Boevé, Department of Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, The Netherlands, and Department of Methodology and Applied Biostatistics, Vrije Universiteit Amsterdam. Rob R. Meijer and Casper J. Albers, Department of Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, The Netherlands; c.j.albers@rug.nl. Hans J. A. Beldhuis, Center for Information Technology, Educational Support and Innovation, University of Groningen. Roel J. Bosker, Department of Educational Science, University of Groningen.

academic dismissal policy on the performance of all first-year students at the university. Here, the natural variation also included factors such as discontinuation of certain courses, changes in lecturers, and educational innovations in other courses. Another layer of complexity occurs when studying performance at multiple academic institutions, possibly even in multiple countries. Then, the grading culture plays a role as well (Fuller, 2013).

Thus, there is not a single definition of "naturally expected variation" that is applicable to all cases. It depends on the context which factors should be included for a particular research question. Conceptually, this is similar to residual variance in analysis of variance (ANOVA) studies: Everything that is not explicitly measured and taken into account by the model is part of the residual variance. Furthermore, as we will explain in the next section, the number of factors that contribute to natural variation is very high, including many factors that are mostly unmeasured in most educational innovation efficacy studies. Thus, a full theoretical approach to taking this naturally expected variation into account is impossible, and a data-driven approach is indispensable.

Thus, in educational innovation settings, the concept of "naturally expected variation" is important to take into account. An important advantage of understanding the extent of "naturally expected variability" of exam scores is that teachers, management, and researchers can anticipate effect sizes necessary to evaluate the success of educational changes. This is especially important in field studies in educational practice, which are often dependent on quasi-experimental designs at best. In this study, we will conduct an analysis on both variation in course grades and pass rates, and we will provide an example of how this information can be used in a research setting.

Prior Research

There is a long history of research into grading throughout all levels of education (Brookhart et al., 2016). In the early 20th century, a lot of research focused on the variability and reliability of grades, mainly in primary and secondary education, but also at college level (Brookhart et al., 2016). Two early extensive literature reviews (Harris, 1931, 1940) summarized that personal factors of students directly related to intelligence, personality, and demographics (age, gender, and family background) have a clear relation to college grades. Also, for example, the type of teaching method, class size, and types of incentives for students might play a role. More recent studies confirm these findings. For instance, college grades are related to class size (Kokkelenberg, Dillon, & Christy, 2007) and lecture attendance (Silvestri, 2003).

In his overviews, Harris (1931, 1940) also provides numerous smaller and less well-known effects on college grades, such as blood pressure, whether students smoke or not, and what proportion of students has liberal political views. More recently, effects on academic performance have been found for social media activity (Junco, Heibergen, & Loken, 2010) and student alcohol consumption (Wolaver, 2007).

Although the focus on college grades, discussed up till now, in research has been necessary and fruitful, research on the "variability" of college grades from a course perspective is lacking. Kostal, Kuncel, and Sackett (2016) found evidence for student grade point average (GPA) inflation between the mid-1990s and 2000s, and argued that instructor leniency must be an important source of the observed grade inflation. Beatty et al. (2015) found that student grades are highly reliable and do not vary much between institutions. There has been some research on grade variability at the primary and secondary level of education. Hollingshead and Childs (2011) showed that there was more variation in grades over time for small schools relative to large schools in Canadian primary education. Wei and Haertel (2011) showed that ignoring the clustering of students in classes within schools led to biased reliability and standard errors of school mean grades. In the context of secondary education, Luyten (1994) showed that there was both systematic variation in mean grades across years for specific subjects as well as systematic variation in mean grades among courses.

The above-mentioned research findings have important implications for the context of understanding the variability of grades in higher education. Given the more limited time, resources, and expertise of teachers to ensure equal exam quality every year, pass rates and mean grades may vary more in higher education as compared to primary and secondary education. On the other hand, the massification of higher education may contribute to smaller standard errors given larger classes compared to primary and secondary education. The clustering of grades is an important factor to take into account as demonstrated by Wei and Haertel (2011). Although research in higher education has often considered student GPA, the clustering of grades within years within courses has not been investigated as far as we know. Similar to secondary education as investigated by Luyten (1994), students in higher education also take different courses taught by different teaching staff. Thus, grades in higher education are also expected to vary among courses as well as within courses across different years.

Although there is little large-scale research on course grades in higher education, course grades are often used in small-scale field studies to investigate various changes or innovations in the learning environment, with sometimes firm conclusions. Therefore, in the present study, we examined the variation in course grades and pass rates in higher education and we illustrate how this information can be used to better compare course mean grades across different years.

Method

Data

Data were obtained from a large university in the Netherlands. The university administration provided assessment records for all first-year courses in bachelor degree programs. The fully anonymized administrative records contained final assessment results from the academic year 2010–2011 through 2015–2016. During these years, the university consisted of nine faculties. This research is classifies as documentary research for which no ethical approval was necessary according to the guidelines of the ethical committee at the university.

Table 1 shows the faculties with the full faculty[1] name and an abridged short description that will be used in the remainder of this article. Table 2 shows the mean grade and overall pass rate per cohort. All courses from the first year of all bachelor degree programs were included. We only used first-year courses because these are obligatory and prerequisite introductory courses for further specializations later in the bachelor degree programs. Using these courses, a good picture could be obtained from the results of complete cohorts.

Table 1. Number of Assessment Observations per Faculty, With Mean Grade (SD) and Overall Pass Rates

Full Faculty Name	Short Name	*N* Assessments	*N* Year Courses	*N* Unique Courses	*N* Unique Students	Mean Grade (*SD*)*	Overall Pass Rate**
Arts	Arts	65,798	1,094	358	9,270	6.74 (.74)	.80
Behavioral & Social Sciences	Social	73,563	427	112	8,155	6.45 (.66)	.77
Economics & Business	Economy	83,952	354	115	9,879	6.25 (.66)	.74
Law	Law	36,953	147	43	5,785	6.18 (.74)	.72
Medical Sciences	Medicine	26,385	221	74	3,945	6.65 (.61)	.80
Philosophy	Philosophy	6,301	110	36	1,388	6.73 (.62)	.83
Science & Engineering	Science	68,209	622	139	6,709	6.67 (.79)	.80
Spatial Sciences	Spatial	11,676	104	30	2,023	6.44 (.65)	.71
Theology & Religious Studies	Theology	2,256	126	33	428	7.10 (.70)	.92
Total		375,093	3,205	940	47,582	6.61 (.74)	.78

*Mean grade (*SD*) is computed as the mean (*SD*) of the mean grades per course.
**Pass rate is computed as the mean of the pass rates per course.

Table 2. Mean Grade and Overall Pass Rate for Each Cohort

Cohort	Mean Grade (*SD*)	Overall Pass Rate
2010	6.66 (.78)	.79
2011	6.57 (.74)	.78
2012	6.66 (.76)	.79
2013	6.57 (.76)	.78
2014	6.60 (.70)	.78
2015	6.61 (.70)	.80

In addition to the full cohorts of enrolled students, second- and third-year students from other bachelor degree programs may also take first-year courses in order to complete a minor. These students were also included in the data. The data we analyzed had the following structure: an anonymous student identifier, a course code, a faculty code, date of examination, and examination result in the form of a grade or pass/fail.

In the data preparation process, after removing empty rows and duplicate records, we selected main course results (thus excluding partial assessment records kept by some faculties), first-attempt results (thus excluding re-sits), and excluded exemption records, resulting in a total of 375,222 assessment records. Subsequently, courses were excluded if only one student participated in the examination, as these courses have no within-course variation ($n = 129$). The final data consisted of a grand total of $N = 375{,}093$ assessment records from 940 unique courses (see Table 1 for further details per faculty). In the appendix, Table A1 shows the distribution of assessment records across faculties and cohorts, and Table A2 shows how many courses per faculty were included and the distribution by number of cohorts per course in the data, with Table A3 showing how many of these courses are unique.

The total number of students in the data equaled $N_S = 40{,}087$, whereas the total number of unique faculty–student combinations was $N_{FS} = 47{,}582$. These numbers imply that some students took first-year program courses in more than one faculty, for example, because they were enrolled in two programs simultaneously. The total number of unique student–year combinations was $N_{SY} = 58{,}612$. This means that some students took courses from first-year bachelor degree programs within the same faculty in different years as a result of, for example, delayed study program due to illness, unforeseen circumstances, double-degree enrollment, and following a minor program from another bachelor program at the same faculty as the main degree of enrollment. Note that only a student's first course enrollment and assessment result were included in the data; thus, there were only unique student–course combinations, and a student–course combination cannot occur more than once in the data.

Measures

The variation in student performance was operationalized by variation in student grades and by whether students passed or failed an exam. As in most continental European countries, a number grading system is employed in the Netherlands. For most courses (specific to each year), 96.8% ($n = 3{,}101$) gave grades on a scale ranging from 1 to 10 where grades of 6 and higher represent a pass. Most degree programs award grades in integers, but sometimes grades are given with decimals. For the present study, all grades were rounded to an integer value. A small part of the courses (specific to each year) 3.2% ($n = 104$) only recorded whether the student passed or failed an exam, thus providing a dichotomous result.

Analyses

Most research on student grades in higher education has focused on student GPA as the main outcome of interest. In order to examine the variation in outcomes across years and among courses in the present study, we focused on course grades. This means that a nested structure was assumed, which is depicted in Figure 1. The illustration of the different nesting structures of interest to the present research on course grades, compared to research on student GPA, illustrate that the same data can be assigned to different levels and that both models are essentially incomplete. In the common perspective of student GPA (Figure 1, left), the lowest level observations are not independent, because each student does not take a new set of courses, but rather some students take the same set of courses (Figure 1, right). Similarly, in the present study, courses in particular years do not all have a unique group of students, and some course years share a common group of students. This observation of nonnestedness is in line with Sun and Pan (2014), who wrote "multilevel data collected in many educational settings are

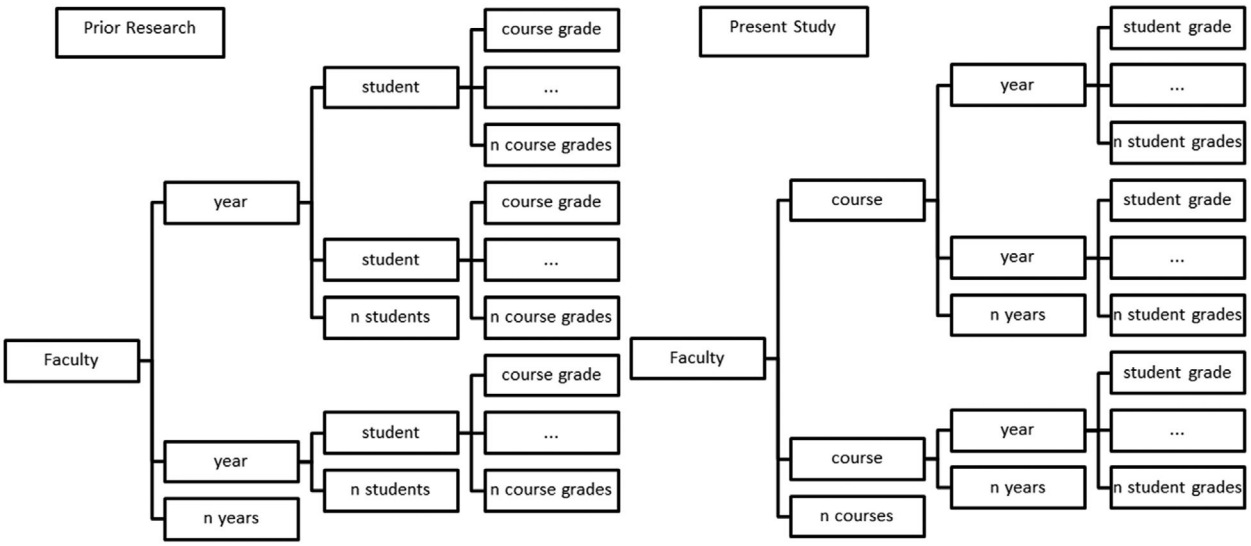

FIGURE 1. Conceptual visualization of the assumed nesting structure in prior research on student GPA (left), and the nesting structure of interest to the research question in the present study (right).

often not purely nested." Meyers and Beretvas (2006) make similar methodological remarks. This complexity in higher education assessment data is an important challenge for researchers, but beyond the scope of the present study to solve, as this would require the programming of new software. In the discussion section, we return to this issue.

Models

We constructed two models: the first model concerned the variation in mean grades and, thus, was applicable to 98.6% of the data. The second model concerned variation in the pass rate. As, obviously, a grade can always be converted into a pass/fail statement, this model is applicable to the full data set.

Model for Mean Grades

The variation in course grade results was examined by estimating an intercept-only multilevel model (Hox, 2010; Snijders & Bosker, 2012) with three levels for student grades as follows:

$$Y_{ijk} = \gamma_{000} + v_{00k} + u_{0jk} + e_{ijk}, \quad (1)$$

where a particular grade Y_{ijk} for student i in year j in course k is modeled by the expected value γ_{000}, with a random error component for the course level (v_{00k}), a random error component for the year level (u_{0jk}), and a residual error component (e_{ijk}). All random components were assumed to be normally distributed around zero.[2] As can be seen from Figure 1, courses are also nested in faculties. However, the number of nine faculties was too small to include as a separate level (Maas & Hox, 2005). In order to explore whether there were differences in mean student performance per faculty, we included faculties as fixed effects with the Faculty of Arts as the reference group. In addition, we examined the proportion of variance at the year and course-level within each faculty by separately estimating the model shown in Equation 1 for each faculty.

The variance decomposition at different levels was investigated in the following way for student grades. First, we examined the total proportion of variance between courses and years as

$$\rho_{course.year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{\sigma^2_{e_{ijk}} + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}, \quad (2)$$

where $\sigma^2_{e_{ijk}}$ denotes the remaining variance in grades at the lowest level, $\sigma^2_{u_{0jk}}$ denotes the variance between years, and $\sigma^2_{v_{00k}}$ represents the variance between courses. The residuals of each level are assumed to have a normal distribution, around 0. Next we examined what proportion of the higher level variation is specific to the year level by

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}. \quad (3)$$

Model for Pass Rates

To model the pass rates, a couple of additional steps were required. To examine variation in pass rates, we set up a multilevel logistic regression model that investigates whether an assessment result was a pass (1) or a fail (0) as

$$\eta_{jk} = \log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \gamma_{000} + v_{00k} + u_{0jk}. \quad (4)$$

Here, π_{jk} is the probability that a student in year j passes course k. This probability consists of an expected value of γ_{000}, a random error component across years (u_{0jk}), and with a random error component across courses (v_{00k}). After the logistic transformation, this can be modeled with the regression model in Equation 4.

After estimating this model, a second model was estimated to explore whether the mean log-odds of passing differed in each faculty. As in the analyses of grades, dummy variables for each faculty were specified with the Faculty of Arts as the reference faculty. In order to explore whether the amount of

course- and year-level variance in log-odds of passing varied across faculties, the intercept only model in Equation 4 was also repeated for each faculty separately.

Log-odds are not straightforward to interpret, but can be transformed back to probabilities using the relation $\pi_{jk} = e^{\eta_{jk}}/(1 + e^{\eta_{jk}})$. In each multilevel model with dichotomous outcomes, the variance of the lowest level is scaled to $\pi^2/3$ (Snijders & Bosker, 2012). This means that in each model for binary outcomes using the logistic link, the residual variance is the same. To examine the variance in log-odds of passing at higher levels, the proportion can be decomposed as

$$\rho_{year} = \frac{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}{\frac{\pi^2}{3} + \sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}, \quad (5)$$

$$\rho_{course} = \frac{\sigma^2_{u_{0jk}}}{\sigma^2_{u_{0jk}} + \sigma^2_{v_{00k}}}. \quad (6)$$

Software

All analyses were conducted in *R* (R Core team, 2017, version 3.4.1), using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015, version 1.13). Full maximum likelihood estimation was used to estimate the model deviance in order to be able to compare the intercept-only model with the model including fixed-effect dummy variables for the different faculties.

Results

To depict the variation in mean course grades, Figure 2 shows the overall mean course grade and the mean course grade for each year within a course for all faculties included in the data.

Course Grades

Table 3 shows the model results for the intercept-only model, and the model with faculty included as a dummy variable in the analyses. Overall, about 17% of the variation in grades can be attributed to systematic variation between courses and years. When adding faculties as a fixed effect by means of dummy variables to the model, there was a statistically significant reduction in the model deviance (Δ deviance = 108, $df = 8$, $p < .001$), implying better model fit, with the proportion of variance being attributed to variation between courses and years being similar to the model without faculty effects. Variation in mean grades between faculties explained about 10% of the variance among courses, which is about 1% of the total variance.

Running a separate intercept-only model for the different faculties shows that the amount of total course and year variation in grades ranged between 11% and 20% (see Table 4). Furthermore, with respect to the higher level amount of variance, Table 4 also showed that the proportion of variance at the year level ranged from 25% to 52%. Because in practice most studies toward the efficacy of an educational innovation take place at the faculty level, we used the intercept-only model for the faculty at which the innovation takes place to compute the benchmark for the natural variation. Another reason to pick this model is that individual faculties can have quite different features (see Table 4), and that the features of Faculty B might not be optimal to infer about an educational innovation at Faculty A.

Pass Rates

Based on the model with the full data, Table 5 indicates that about 40% of the variance in the log-odds of passing was at the year level and course level. Approximately 23% of the higher-level variance was due to differences between years within courses. The inclusion of fixed effects for faculties yielded a significant better model fit (Δ deviance = 62, $df = 8$, $p < .001$). Table 6 shows that there is considerable variability among faculties in the amount of variance in log-odds at the year level and course level, with estimates ranging from 22% to 74%. Furthermore, the relative amount of variance at the year level within a course rather than among courses also varied considerably, from 5% to 70%. It is important to note that these percentages of variability at the log-odds level do not translate easily to percentages at the pass or fail level, which will be made clear in the following.

Application

Consider the following scenario, with intentionally simplified numbers. A course instructor is interested in implementing a new teaching method. It is not possible to do a RCT and the instructor would like to compare the results of the previous year, that is, the year prior to the implementation of the new teaching method, with that of the current year, now that the changes have been implemented. In both years, $n = 50$ students participate in the course, and the GPA is 6.00 in the first year and 6.50 in the second (on the 1–10 scale). In both years, the standard deviation of the grades is 1.00. A standard t-test shows that the increase in GPA is highly significant ($t = 2.50$, $df = 98$, one-sided $p = .007$). Concluding that, thus, the new teaching method is beneficial is misleading, as the regular year-to-year variations are not taken into account. To infer a significant increase in GPA after an educational intervention, the increase in GPA should not just be significantly above zero, but significantly above regular values obtained from year-to-year variation.

The variance partitioning of grades and year variation in the present study can be informative. Based on the estimated proportion of variance across years, a course instructor can estimate the 95% confidence interval around the difference between two cohort mean grades as follows:

$$0 \pm t^* \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \left(\sigma^2_{\text{residual}}\right) + \sigma^2_{\text{year}}}, \quad (7)$$

where n_1 and n_2 are the number of students participating in the course in each years and t^* is the critical t-value with $n_1 + n_2 - 2$ degrees of freedom. The course-level variance is excluded here because the result in both years is for the same course. For a random course, the year-level variance component of the overall model can be used based on the intercept-only model. It is also possible to use a faculty-specific variance component if the faculty is known. Figure 3 shows the 95% confidence interval around the mean grade for different possible numbers of students in each cohort, based on the estimated variance components of the overall model. From this figure, it is clear that an increase of .5 in GPA for a

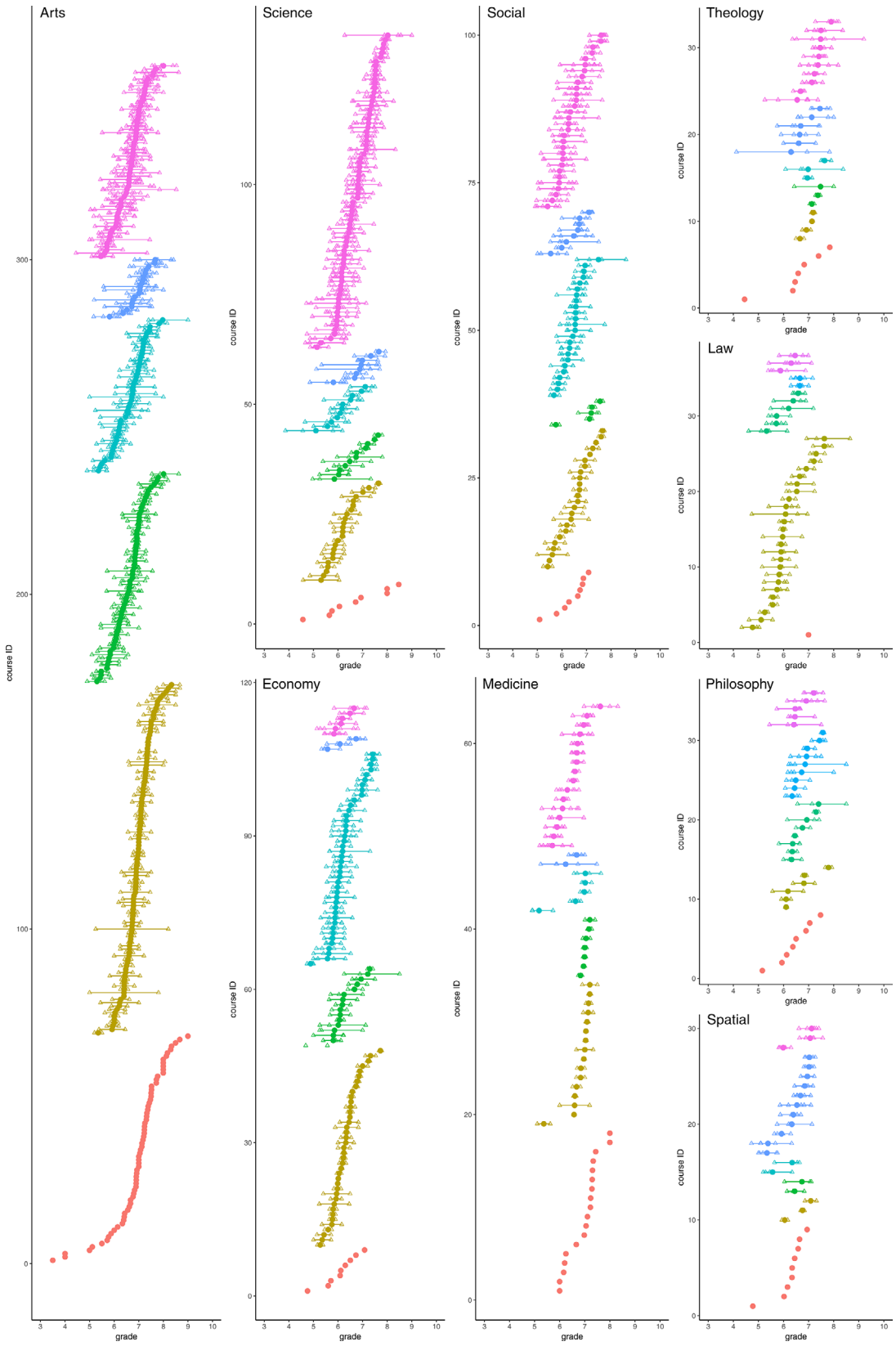

FIGURE 2. The variation of mean course grades within and between each course in each faculty included in the data with colors indicating the different number of cohorts for each course (within each faculty, from top [six cohorts] to bottom [one cohort]). Each horizontal line represents the distance between the lowest mean year grade and highest mean year grade for each course, with triangles representing mean grade in each year, and closed circles the mean grade over cohorts for each course. [Color figure can be viewed at wileyonlinelibrary.com]

Table 3. Estimates of the Variance Components, and Model Deviance for Course Grades

	Intercept-Only Model	Model Including Faculty Fixed Effect
Fixed Effects		
Intercept γ_{000}	6.59 (*.02*)	6.76 (.03)
$D_{Theology}$		.25 (.11)
D_{Law}		–.60 (.10)
$D_{Medicine}$		–.01 (.08)
$D_{Science}$		–.16 (.06)
$D_{Economy}$		–.51 (.06)
D_{Social}		–.30 (.07)
$D_{Philosophy}$		–.09 (.11)
$D_{Spatial}$		–.36 (.11)
Random Effects		
Grades $\sigma^2_{e_{ijk}}$	2.27	2.27
Years $\sigma^2_{u_{0jk}}$	.15	.14
Courses $\sigma^2_{v_{00k}}$	.32	.28
Deviance	1,294,263	1,294,156
Δ Deviance		108
$\rho_{course:year}$	.17	.16
ρ_{year}	.31	.34

Table 4. Variance Partition of Grades at the Different Levels for Each Faculty

Faculty	Residual Variance	Year Variance	Course Variance	$\rho_{course:year}$	ρ_{year}
Arts	1.92	.16	.26	.18	.38
Economy	2.59	.12	.25	.13	.33
Law	2.87	.21	.33	.16	.38
Medicine	1.34	.09	.24	.19	.29
Philosophy	2.31	.14	.13	.11	.52
Science	2.36	.15	.44	.20	.25
Social	2.23	.15	.26	.16	.37
Spatial	1.50	.11	.27	.20	.30
Theology	1.44	.21	.14	.20	.41

Table 5. Results of the Random Intercept Models on the Log-Odds of Passing

	Intercept Only	Model With Faculty as Fixed Effect
Fixed effects		
Intercept γ_{000}	1.80 (*.05*)	1.82 (.07)
$D_{Theology}$		1.73 (.29)
D_{Law}		–.40 (.23)
$D_{Medicine}$		.21 (.18)
$D_{Science}$		.11 (.14)
$D_{Economy}$		–.44 (.14)
D_{Social}		–.05 (.15)
$D_{Philosophy}$		.04 (.24)
$D_{Spatial}$		–.59 (.27)
Random effects		
Years $\sigma^2_{u_{0jk}}$	.5086	.5106
Courses $\sigma^2_{v_{00k}}$	1.7315	1.6206
Deviance	366,947	366,885
Δ Deviance		62
$\rho_{year + course}$	.41	.39
ρ_{year}	.23	.24

Table 6. Coefficients for the Random Intercept Models on the Grades and Log-Odds of Passing for Each Faculty

Faculty	Year Variance	Course Variance	$\rho_{year + course}$	ρ_{year}
Arts	.47	.84	.28	.36
Economy	.21	1.36	.32	.13
Law	.20	3.45	.53	.05
Medicine	.14	2.24	.42	.06
Philosophy	.24	.70	.22	.25
Science	1.01	2.91	.54	.26
Social	.54	1.95	.43	.22
Spatial	1.40	.60	.38	.70
Theology	2.60	6.94	.74	.27

course with 50 students per year is nonsignificant. For larger courses, for example, with 200 students per year, a .5 increase would be a significant effect.

Similarly, Equation 7 can also be used to estimate the 95% confidence interval around the log-odds of passing. In contrast to the application of mean grades, this is, however, dependent on the intercept (i.e., the log-odds of the average pass grade), while the application for grades is equivalent regardless of the mean expected grade. Figure 4 shows the same 95% confidence interval after transforming the log-odds interval back to the probability of passing. We assume that a teacher observes that the original cohort had a pass rate of .86, and observes a pass rate of .90 in the course with the new lecture method. Then, Figure 4 shows that you need at least 150 students per year for this difference to be significant at the 5% level.

To illustrate how the confidence interval of the log-odds varies depending on the expected intercept, Figure 4 shows the interval for different possible numbers of students given three different intercepts, based on the quantiles of pass rates in the present data. This figure shows that whether a certain increase from year 1 to year 2 in pass rate is significant depends on the pass rate of year 1. For instance, in a course with 100 students, a 5 percentage point increase from 60% to 65% is not significant, whereas the same increase from 90% to 95% would be. In general, for pass rates closer to 1 (or to 0), a smaller increase in pass rate can be more significant than for pass rates closer to 50%.

Example: The Flipped Classroom

To highlight the consequences of not taking natural variation into account, we considered studies on the flipped classroom. One of the main intended goals of this now popular intervention is to increase student performance. Five recent papers (Liebert, Lin, Mazer, Bereknyei, & Lau, 2016; Mason, Shuman, & Cook, 2013; Pierce & Fox, 2012; Street, Gilliland, McNeil, & Royal, 2015; Tune, Sturek, & Basile, 2013) studied the effect of the flipped classroom on, among others, student performance using a two-cohort design: the first cohort received "traditional" instruction for a course and the second cohort received some form of flipped classroom instruction. Student performance of both cohorts was then compared using a t-test.

For these studies, we recalculated the p-values, taking natural cohort variation into account. The paper by Pierce and Fox (2012) unfortunately provided insufficient information

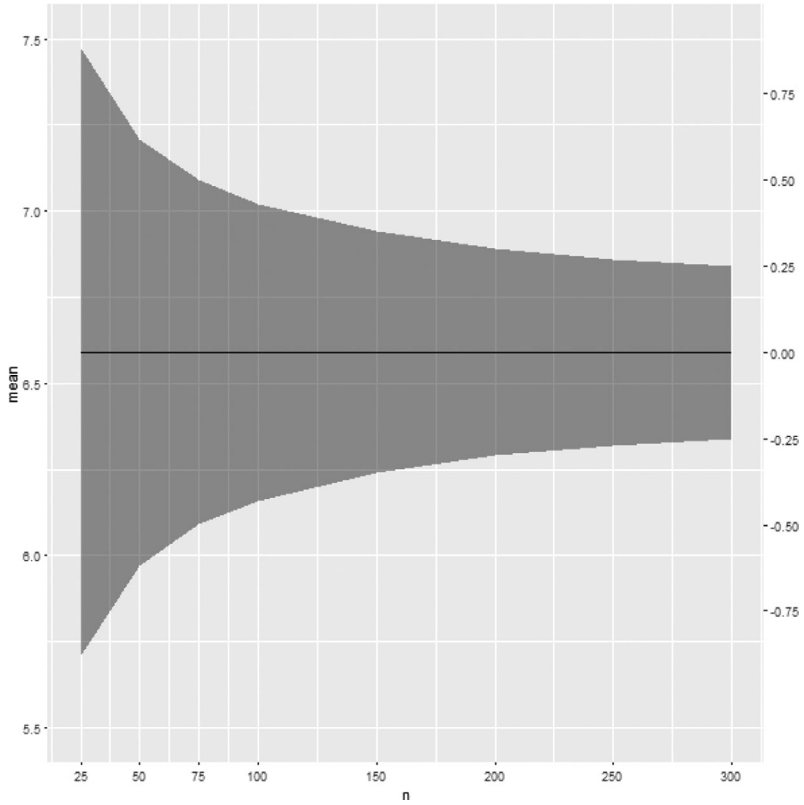

FIGURE 3. The 95% confidence interval around the predicted mean grade when $n_1 = n_2$. The horizontal line is placed at the observed mean grade in the data set (left vertical axis), but the width of the confidence interval would be the same when placed around another mean value. The right vertical axis displays the deviation from this mean value.

Table 7. Comparison of p-Values for Four Studies on the Flipped Classroom

	Cohort 1 (Traditional)		Cohort 2 (Flipped)		*p*-Value		
Study	**Sample Size**	**Mean (*SD*)**	**Sample Size**	**Mean (*SD*)**	**Reported**	**Adjusted (half)**	**Adjusted**
Tune et al. (2013)	14	6.76 (1.91)	13	7.98 (1.27)	.03	.044	.055
Mason et al. (2013)	20	8.34 (.70)	20	9.09 (.71)	.002	.040	.101
Street et al. (2015)	180	8.5 (.70)	180	8.7 (.70)	.026	.241	.306
Liebert et al. (2016)	92	7.57 (.82)	89	7.48 (.81)	.28	.371	.403

Note: Reported *p*-values are the ones as reported in the literature: the adjusted ones take natural variation into account, at either half the level or the full level of natural variation observed at the university studied in this manuscript.
The studies by Tune et al. (2013) and Mason et al. (2013) reported scores on a scale to 10: for the other two studies the values have been rescaled to this scale for comparison purposes. Note that in the case of Liebert et al. (2016), the effect is in the opposite direction (i.e., traditional instruction outperforms the flipped instruction). The values from Tune et al. (2013) were obtained by manually reading off figures in that paper. In the study by Mason et al. (2013), 22 performance tests were conducted without correcting for multiple testing. As this statistical discrepancy lies outside the scope of this paper, we did not correct for it and simply reported the unadjusted *p*-value.

(e.g., no information of the sample size of the control cohort) to be included in the computations. For the other four studies, Table 7 displays the key characteristics such as sample size and GPA increase. It also shows the *p*-value corresponding to the intervention according to the original authors, as well as when taking natural variation into account. Other universities will not have exactly the same level of natural variation as the Dutch university studied here, but will certainly have more natural variation than the zero level implicitly assumed by the original authors. However, it could be that—for some unknown reason—variability at the Dutch university studied here is considerably larger than at other institutions. In that case, the adjusted *p*-values in Table 7 will be too high. For this reason, we also computed the adjusted *p*-values in case the natural variation is only half the size of that at the Dutch university studied.

We have no reasons to expect that grade fluctuations at this Dutch university would differ strongly from other universities, yet also no evidence that the fluctuations are very similar. The goal of this example is not to compute the exact value of the adjusted *p*-value because then we need more information on the natural variation at the institutions studied. Thus, which of the other two columns with adjusted *p*-values is more correct is a matter of future research, but both give the same message: After adjusting for natural variation between cohorts, the evidence for a beneficial effect of flipped classrooms on exam performance is much less clear. Not taking this variation into account leads to too many false positive findings. It is completely unreasonable to expect zero natural variation, and thus the important message is that originally reported *p*-values in Table 7 will certainly be too small.

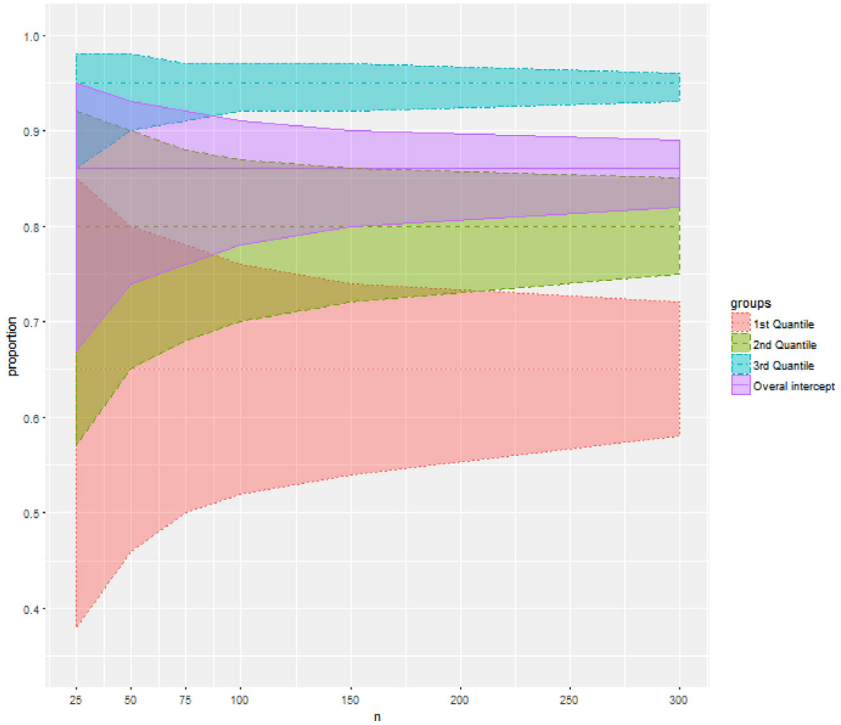

FIGURE 4. The 95% confidence interval of the probability of passing based on the overall model intercept (mean of .86), and the quantiles of mean pass rates, when $n_1 = n_2$. [Color figure can be viewed at wileyonlinelibrary.com]

Discussion

The aim of the present study was to explore the extent to which assessment results, both in terms of grades and in terms of passing, vary between years within courses and the extent to which they vary across courses within different faculties. As there are too many factors playing a role in the natural variation, either by random fluctuation or by some (unknown) trend over time, it is unfeasible to measure and model them all, especially when it concerns variation over years. Therefore, in this study, we used a data-driven approach for capturing the natural variation.

We discussed that, depending on the research question, some concepts either are or are not part of the natural variation. Disregarding the different disciplines of the different courses, the present study found that about 17% of the variation in grades was at the year level and course level. Of this variation, about 30% was due to variation within courses across different years, while the remaining 70% was due to systematic variation among courses. Despite the high reliability of student GPA as demonstrated by Beatty et al. (2015), the present study showed that year-over-year variations in grades are considerable.

When examining the log-odds of whether an assessment result was a pass or a fail, we found that approximately 40% of the variance was at the year level and course level, with 25% of this variation across different years within courses and 75% among different courses overall. When accounting for different disciplines (faculties) in the data, the amount of variation among courses decreased slightly from 17% to 16% in terms of course grades, and from 41% to 39% in the log-odds of passing.

In line with the findings of Luyten (1994) in the context of secondary education, the present study found that the proportion of course-level variation was larger than the variation within courses across years. However, exploring discipline-specific differences in the amount of variation at the course level and year level revealed substantial differences among faculties. The overall amount of higher level variation varied from 11% to 20% concerning grades, and ranged from 22% to 72% for the log-odds of passing. Regarding the higher level variance, the proportion of variance across years ranged from 25% to 52% for grades, and for the log-odds of passing, the proportion of variance across years relative to the higher-level variance ranged from 5% to 70%.

Take-Home Messages

There are some important take-home messages from our findings as was illustrated in the application section. In the educational literature, innovations are often judged effective based on a direct comparison of two cohorts, without taking the "naturally expected variation" into account. That is, these studies are treated as RCTs, whereas they are, at best, quasi-experimental field studies. Disregarding the general fluctuation in course grades over time leads to a severe increase of false positives as innovations may be incorrectly labeled as effective. Although for a course with 50 students, a difference in grade mean of half a standard deviation before and after intervention would be considered highly significant ($p = .07$) when disregarding this variation, the difference actually is nonsignificant at the $\alpha = .05$ threshold. At least 75 students per year are needed to get the p-value below .05, and many more to get it at the value .07 that would be reported by those ignoring the natural variation.

In line with the findings of Hollingshead and Childs (2011), as the number of students increases the uncertainty around both the mean course grade and pass rates decreases. This study demonstrated that, even with large sample sizes,

conclusions about cohort differences should be taken with caution. For instance, for a large course, with 300 students per year, an increase in pass rate from, say, 65% to 70% is not significant. When ignoring the natural variation, this difference would be highly significant.

As evaluations of educational innovations ignore this natural variation, it is to be expected that the number of false positive findings is very large. A practical recommendation to avoid this is as follows. Based on the number of students in a course, one can use Figure 3 to find the value δ, which is the maximum value of the difference in mean grades in two consecutive years, $m_2 - m_1$, which would be nonsignificant; for instance, with $n = 50$, $\delta = .62$. Rather than testing for a significant difference between both means (H_0: $\mu_2 - \mu_1 = 0$, with the standard t-test), one can then test whether the difference between both means is significantly larger than δ or not. For this, one can use equivalence tests (Lakens, 2017; Schuirmann, 1987). To claim a successful educational innovation, the difference between grade means should significantly exceed δ, rather than just significantly exceed 0. When the interest lies in the pass rate rather than the grade mean, a similar approach can be employed using Figure 4.

Furthermore, it would be interesting to recompute the p-values for (popular) educational innovations that claimed to have been significant, but this time taking the natural variation into account. Alternatively, a replication study of the innovation can shed new light onto the reliability of certain innovations.

Limitations

The present study was focused on assessment in higher education. As always in data analysis, not all potentially relevant variables were measured. Some faculties offer multiple bachelor degree programs, and there may be systematic variation among bachelor programs within the same faculty. Because our data set did not provide information on which bachelor program a course belonged to, we could not take this level into account in our analyses. Also, the effect of individual teachers could not be taken into the model as this information was not part of the data set.

Another limitation in the present study was that it was unknown to what extent courses were taught in the same way or by the same teachers in different years. There were some major education innovations, but these "partly new" courses could not be identified in our data set. The variance across years, however, likely does include teacher experimentation with new technology or assessment methods. Also, note that the average grade did not increase significantly over the years (Table 2).

The main limitation of this study concerns the generalizability of the results. In the present study, we examined the grades and pass rates of first-year courses in higher education at a single, large university in the Netherlands. Given the large amount of information, the estimated variance components could be informative for other institutions, especially those using a number grading scale. Although it is unknown to what extent the numerical findings in this study are representative for other universities, it is obvious that also at other places a considerable part of grade variation can be labeled as "natural variation." Thus, the message that many "significant" findings in assessing educational interventions are actually false positives holds, but further research is needed to assess "how" many of these findings are false.

In future studies, higher education institutes can employ the model introduced by us on their own assessment records. If they find more natural variation at their institute than we did in our study, an even larger grade increase is required for a successful intervention. Reversely, with less natural variation, smaller increases can be labeled as successful.

The present study demonstrated that assigning observations to different levels is sometimes not straightforward (see, e.g., Hox, 2010). For research on student grades in higher education, the focus has often been on the student with the interest in explaining why individuals differ in their achievement level; here, the focus was on how courses differed in achievement across different years. A cross-classified model would allow for a combination of both types of nesting in the same model (Hox, 2010), but this approach was currently not feasible for us. Specialized software such MLWin cannot handle data sets as large as ours, which is why we used *R*. For *R*, currently no packages exist for such cross-classified models.

The design and model used in this paper constituted a violation of the independence assumption, as the data set contains multiple records per student. This was unavoidable, as no software package currently provides better alternatives for data sets of this size. The development of a completely new statistical model was beyond the scope of this paper. Furthermore, as the results were not intended to be used as error-free benchmark for other studies, some bias was considered acceptable. In Boevé (2018), an ad hoc approach to this data set is presented. Here, 25 samples were drawn from the original data set with only a single measurement per individual. This removed the independence violation of our model but introduced a new one: Students with only a few exams were overrepresented in this data set. Additional pragmatic steps were taken to minimize this bias. This approach led to somewhat larger variance estimates than those presented in Table 3, but it is unknown which approach is better. For this, future methodological research is needed. That most multilevel data in educational settings are not fully nested, and that this is often overlooked by educational researchers, was already observed by Sun and Pan (2014). Meyers and Beretvas (2006) argue that this nonnestedness and the resulting violation of the independence assumption could lead to an inflation of false positives (p. 493). That would imply that the consequences of natural variation are even larger than suggested in Tables 3–6, and that the confidence intervals in Figures 3 and 4 should be even wider.

Conclusion

The goal of this study was three fold: (i) introducing a model for assessing "natural variability" in grades in higher education, (ii) estimating the parameters in this model based on a large ($n = 375{,}093$) data set from a single university, and (iii) illustrating that the consequences of the common practice of ignoring this natural variation in studying whether an educational intervention are very severe, yielding highly inflated results.

The assessment records of higher education institutes contain valuable information when examined from a course perspective rather than from a student perspective. Understanding the variation in course results across years can help teachers and institutions to evaluate the impact of innovations at a cohort level, while reducing the risk of false positives when grades between two subsequent cohorts are compared.

Author contributions

AJB: data analysis and interpretation, and drafting the article; CJA: conception, supervised data analysis, and drafting the article; RRM: critical revision; HJAB: critical revision and supply of the data; RJB: critical revision.

Notes

[1]Note that the use of the word *faculty* does not refer to academic staff, but to independent organizational units within the university roughly organized by academic discipline.

[2]As the scores Y_{ijk} are on an integer scale from 1 to 10, this assumption obviously cannot hold exactly. The main focus is to estimate the variability of the different levels in the model, not individual grades. Furthermore, the central limit theorem states that with samples this large, the consequence of violations of this assumption is extremely small (cf. Ernst & Albers, 2017).

References

Albers, C. J., Vermue, C., de Wolff, T., & Beldhuis, H. (2018). Model-based academic dismissal policies: A case-study from the Netherlands. *PsyArxiv*. https://doi.org/10.31234/osf.io/6a9cz

Alexander, F. K. (2000). The changing face of accountability: Monitoring and assessing institutional performance in higher education. *The Journal of Higher Education*, *71*, 411–431. https://doi.org/10.2307/2649146

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, A. S., Walmsley, P. T., Sackett, P. R., & Kuncel, N. R. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, *34*(4), 31–40. https://doi.org/10.1111/emip.12096

Boevé, A. J. (2018). *Implementing assessment innovations in higher education*. Groningen, The Netherlands: University of Groningen.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., ... Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*, 803–848. https://doi.org/10.3102/0034654316672069

Carey, T. A., & Stiles, W. B. (2015). Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology and Psychotherapy*, *23*, 87–95. https://doi.org/10.1002/cpp.1942

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—A systematic review of common misconceptions. *PeerJ*, *5*, e3323. https://doi.org/10.7717/peerj.3323

Fuller, M. B. (2013). An empirical study of cultures of assessment in higher education. *NCPEA Education Leadership Review*, *14*(1), 20–27.

Harris, D. (1931). The relation to college grades of some factors other than intelligence. *Archives of Psychology*, *20*, 131.

Harris, D. (1940) Factors affecting college grades: A review of the literature, 1930–1937. *Psychological Bulletin*, *37*, 125–166. https://doi.org/10.1037/h0055365

Hollingshead, L., & Childs, R. A. (2011). Reporting the percentage of students above a cut-score: The effect of group size. *Educational Measurement: Issues and Practice*, *30*(1), 36–43. https://doi.org/10.1111/j.1745-3992.2010.00198.x

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Oxford, UK: Routledge.

Junco, R., Heibergen, G., & Loken, E. (2010). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning*, *27*, 119–132. https://doi.org/10.1111/j.1365-2729.2010.00387.x

Kokkelenberg, E. C., Dillon, M., & Christy, S. M. (2007). The effects of class size on student grades at a public university. *Economics of Education Review*, *27*, 221–233. https://doi.org/10.1016/j.econedurev.2006.09.011

Kostal, J. W., Kuncel, N. R., & Sackett, P. R. (2016). Grade inflation marches on: Grade increases from the 1990s to 2000s. *Educational Measurement: Issues and Practice*, *35*(1), 11–20. https://doi.org/10.1111/emip.12077

Lakens, D. (2017). Equivalence tests: A practical primer for *t*-tests, correlations and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. https://doi.org/10.1177/1948550617697177

Liebert, C. A., Lin, D. T., Mazer, L. M., Bereknyei, S., & Lau, J. N. (2016). Effectiveness of the surgery core clerkship flipped classroom: A prospective cohort trial. *American Journal of Surgery*, *211*, 451–457. https://doi.org/10.1016/j.amjsurg.2015.10.004

Liu, O. L. (2011). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Educational Measurement: Issues and Practice*, *30*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2011.00206.x

Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research*, *21*, 197–216. https://doi.org/10.1016/0883-0355(94)90032-9

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86

Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. *IEEE Transactions on Education*, *56*, 430–435. https://doi.org/10.1109/TE.2013.2249066

Meyers, J. L., & Beretvas, N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*, 473–497. https://doi.org/10.1207/s15327906mbr4104_3

Pierce, R., & Fox, J. (2012). Vodcasts and active-learning exercises in a "flipped classroom" of a renal pharacotherapy module. *American Journal of Pharmaceutical Education*, *76*, 196.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680. https://doi.org/10.1007/BF01068419

Silvestri, L. (2003). The effect of attendance on undergraduate methods course grades. *Education*, *123*, 483–486.

Snijders, T. A. B., & Bosker, R. R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Street, S. E., Gilliland, K. O., McNeil, C., & Royal, K. (2015). The flipped classroom improved medical student performance and satisfaction in a pre-clinical physiology course. *Medical Science Education*, *25*, 35–42. https://doi.org/10.1007/s40670-014-0092-4

Sun, S., & Pan, W. (2014). A methodological review of statistical methods for handling multilevel non-nested longitudinal data in educational research. *International Journal of Research and Method in Education*, *37*, 285–308. https://doi.org/10.1080/1743727X.2014.885012

Tune, J. D., Sturek, M., & Basile, D. P. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. *Advances in Physiological Education*, *37*, 316–320. https://doi.org/10.1152/advan.00091.2013

Wei, X. & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice*, *30*(1), 13–22. https://doi.org/10.1111/j.1745-3992.2010.00196.x

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., ... Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*, 1359–1366. https://doi.org/10.2105/AJPH.2007.124446

Wolaver, A. M. (2007). Does drinking affect grades more for women? Gender differences in the effects of heavy episodic drinking in college. *American Economist*, *51*, https://doi.org/10.1177/056943450705100211

Appendix

Table A1. Number of Assessment Observations per Faculty in Each Year

Faculty	2010	2011	2012	2013	2014	2015	Total
Theology	394	371	469	358	305	359	2,256
Law	7,683	7,391	7,281	4,703	5,051	4,844	36,953
Medicine	4,993	4,588	4,445	4,831	3,976	3,552	26,385
Science	10,377	10,735	10,991	12,053	11,818	12,235	68,209
Arts	11,923	11,021	10,686	11,325	10,477	10,366	65,798
Economy	15,514	14,461	13,495	14,392	13,954	12,136	83,952
Social	13,758	13,552	13,312	12,808	11,132	9,001	73,563
Philosophy	1,173	1,261	1,311	956	755	845	6,301
Spatial	2,337	2,108	1,935	1,890	2,034	1,372	11,676
Total	68,152	65,488	63,925	63,316	59,502	54,710	375,093

Table A2. Number of Courses per Faculty in Each Year

	2010	2011	2012	2013	2014	2015	Total
Theology	17	20	22	22	23	22	126
Law	28	28	28	26	20	17	147
Medicine	32	36	36	40	40	37	221
Science	101	104	106	101	106	104	622
Arts	208	206	171	168	174	167	1,094
Economy	66	70	67	51	50	50	354
Social	69	73	74	75	69	67	427
Philosophy	21	20	19	24	13	13	110
Spatial	18	18	19	18	18	13	104
Total	560	575	542	525	513	490	3,205

Table A3. Number of Unique Courses per Faculty, and Distribution of Courses by Number of Cohorts Included in the Data

Faculty	1 cohort	2 cohorts	3 cohorts	4 cohorts	5 cohorts	6 cohorts
Theology	7	4	3	3	6	10
Law	1	0	31	5	3	3
Medicine	22	20	7	5	2	18
Science	10	24	11	13	7	74
Arts	67	105	62	47	19	58
Economy	9	39	16	42	3	6
Social	8	32	6	25	9	32
Philosophy	8	6	8	9	0	5
Spatial	8	4	1	3	11	3
Total	140	234	145	152	60	209

